

A decorative border made of film strips in cyan and purple colors, framing the central text area.

Movie Analysis

To -> Mrinal Das Sir

Shivendr Srivastava (142402019)
Sudhin S (142402022)

PART – 1

**[SHIVENDR
SRIVASTAVA]**

Pre - Processing

Pandas Regex

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	
1	id	imdb_id	popularity	budget	revenue	original_title	cast	homepage	director	tagline	keywords	overview	runtime	genres	production	release_date	vote_count	vote_average	release_year	budget_adj	revenue_adj	
2	135397	tt036961	32.9858	1.5E+08	1.5E+08	Jurassic V	Chris Pratt	http://ww	Colin Trev	The park i	monster	Twenty-tv	124	Action	Ac Universal	#####	5562	6.5	2015	1.4E+08	1.4E+09	
3	76341	tt139219	28.4199	1.5E+08	3.8E+08	Mad Max: Tom Hard	http://ww	George M	What a Lc future	ch An apocal			120	Action	Ac Village Ro	5/13/15	6185	7.1	2015	1.4E+08	3.5E+08	
4	262500	tt290844	13.1125	1.1E+08	3E+08	Insurgent	Shailene	http://ww	Robert Sc	One Choic based on	Beatrice f		119	Adventur	Summit Ei	3/18/15	2480	6.3	2015	1E+08	2.7E+08	
5	140607	tt248849	11.1731	2E+08	2.1E+09	Star Wars	Harrison	http://ww	J.J. Abram	Every gen	android	s Thirty yea	136	Action	Ac Lucasfilm	12/15/15	5292	7.5	2015	1.8E+08	1.9E+09	
6	168259	tt282085	9.33501	1.9E+08	1.5E+09	Furious 7	Vin Dies	http://ww	James W	Vengeanc	car race	s Deckard S	137	Action	Cr Universal	#####	2947	7.3	2015	1.7E+08	1.4E+09	
7	281957	tt166320	9.1107	1.4E+08	5.3E+08	The Rever	Leonardo	http://ww	Alejandro	n. One w	father-so	In the 182	156	Western	Regency E	12/25/15	3929	7.2	2015	1.2E+08	4.9E+08	
8	87101	tt134013	8.65436	1.6E+08	4.4E+08	Terminat	Arnold Sci	http://ww	Alan Tayl	Reset the	saving the	The year i	125	Science F	Paramou	6/23/15	2598	5.8	2015	1.4E+08	4.1E+08	
9	286217	tt365938	7.6674	1.1E+08	6E+08	The Marti	Matt Dam	http://ww	Ridley Scc	Bring Him	based on	During a r	141	Drama	Ac Twentiet	9/30/15	4572	7.6	2015	9.9E+07	5.5E+08	
10	211672	tt229364	7.40416	7.4E+07	1.2E+09	Minions	Sandra Bu	http://ww	Kyle Bald	Before Gr	assistant	Minions S	91	Family	Ac Universal	6/17/15	2893	6.5	2015	6.8E+07	1.1E+09	
11	150540	tt209667	6.3268	1.8E+08	8.5E+08	Inside Ou	Amy Poeh	http://mc	Pete Doct	Meet the	dream	ca Growing u	94	Comedy	Ac Walt Disn	#####	3935	8	2015	1.6E+08	7.9E+08	
12	206647	tt237971	6.20028	2.5E+08	8.8E+08	Spectre	Daniel Cri	http://ww	Sam Meni	A Plan No	spy	base A cryptic r	148	Action	Ac Columbia	10/26/15	3254	6.2	2015	2.3E+08	8.1E+08	
13	76757	tt161766	6.18937	1.8E+08	1.8E+08	Jupiter As	Mila Kun	http://ww	Lana Wac	Expand y	c jupiter	sq In a unive	124	Science F	i Village Ro	#####	1937	5.2	2015	1.6E+08	1.7E+08	
14	264660	tt047075	6.11885	1.5E+07	3.7E+07	Ex Machi	Domhnall	http://exr	Alex Garl	There is n	dancing	s Caleb, a 2	108	Drama	Ac DNA Films	1/21/15	2854	7.6	2015	1.4E+07	3.4E+07	
15	257344	tt212012	5.985	8.8E+07	2.4E+08	Pixels	Adam San	http://ww	Chris Col	Game On	video ga	n Video ga	105	Action	Ac Columbia	7/16/15	1575	5.8	2015	8.1E+07	2.2E+08	
16	99861	tt239542	5.94493	2.8E+08	1.4E+09	Avengers: Robert Dc	http://ma	Joss Whe	A New Age	marvel co	When Tor		141	Action	Ac Marvel St	4/22/15	4304	7.4	2015	2.6E+08	1.3E+09	
17	273248	tt346025	5.8984	4.4E+07	1.6E+08	The Hate	Samuel L.	http://the	Quentin T	No one co	bounty hu	Bounty hu	167	Crime	Dr Double E	12/25/15	2389	7.4	2015	4E+07	1.4E+08	
18	260346	tt244604	5.74976	4.8E+07	3.3E+08	Taken 3	Liam Nee	http://ww	Olivier M	It Ends He	revenge	s Ex-govern	109	Crime	Ac Twentiet	#####	1578	6.1	2015	4.4E+07	3E+08	
19	102899	tt047897	5.57318	1.3E+08	5.2E+08	Ant-Man	Paul Ruds	http://ma	Peyton R	Heroes Di	marvel co	Armed wi	115	Science F	i Marvel St	7/14/15	3779	7	2015	1.2E+08	4.8E+08	
20	150639	tt166119	5.55682	9.5E+07	5.4E+08	Cinderell	Lily James	Cate Bla	Kenneth	Midnight	cinderell	When her	112	Romance	Walt Disn	#####	1495	6.8	2015	8.7E+07	5E+08	
21	131634	tt195126	5.47696	1.6E+08	6.5E+08	The Hung	Jennifer L	http://ww	Francis L	The fire w	revolution	With the	136	War	Ac Studio Ba	11/18/15	2380	6.5	2015	1.5E+08	6E+08	
22	158852	tt196441	5.46214	1.9E+08	2.1E+08	Tomorrow	Britt Robe	http://mc	Brad Bird	Imagine a	inventor	s Bound by	130	Action	Ac Walt Disn	5/19/15	1899	6.2	2015	1.7E+08	1.9E+08	
23	307081	tt179868	5.33706	3E+07	9.2E+07	Southpaw	Jake Gyllen	haal	Rai Antoine	F Believe	in sport	s Billy	The	123	Action	Dr Escape Ar	6/15/15	1386	7.3	2015	2.8E+07	8.4E+07
24	254128	tt212635	4.90783	1.1E+08	4.7E+08	San Andre	Dwayne J	http://ww	Brad Peyt	A rescue	s californi	a In the afte	114	Action	Dr New Line	5/27/15	2060	6.1	2015	1E+08	4.3E+08	
25	216015	tt232244	4.7104	4E+07	5.7E+08	Fifty Shad	Dakota Jo	https://w	Sam Tayl	Are you c	i based on	When col	125	Drama	Ac Focus Fea	#####	1865	5.3	2015	3.7E+07	5.2E+08	
26	318846	tt159636	4.64805	2.8E+07	1.3E+08	The Big Sh	Christian	http://ww	Adam McI	This is a	ti bank	s fra	The men v	130	Comedy	Ac Paramou	#####	1545	7.3	2015	2.6E+07	1.2E+08
27	177677	tt238124	4.56671	1.5E+08	6.8E+08	Mission: I	Tom Cruis	http://ww	Christoph	Desperat	spy	sequ Ethan and	131	Action	Ac Paramou	7/23/15	2349	7.1	2015	1.4E+08	6.3E+08	
28	214756	tt263227	4.56455	6.8E+07	2.2E+08	Ted 2	Mark Wahlberg	Sar	Seth Mac	Ted is Co	room	ba Newbu	115	Comedy	Ac Universal	6/25/15	1666	6.3	2015	6.3E+07	7E+08	

```
print(movies_df.isnull().sum())
```

id	0
imdb_id	10
popularity	0
budget	0
revenue	0
original_title	0
cast	76
homepage	7930
director	44
tagline	2824
keywords	1493
overview	4
runtime	0
genres	23
production_companies	1030
release_date	0
vote_count	0
vote_average	0
release_year	0
budget_adj	0
revenue_adj	0
dtype: int64	

Null Values

Total 7 columns of them

Replaced them with
“N/A”

```
print(movies_df.isnull().sum())
```

id	0
imdb_id	0
popularity	0
budget	0
revenue	0
original_title	0
cast	0
homepage	0
director	0
tagline	0
keywords	0
overview	0
runtime	0
genres	0
production_companies	0
release_date	0
vote_count	0
vote_average	0
release_year	0
budget_adj	0
revenue_adj	0
dtype: int64	

Multiple type of Date Formats

	release_date	date_format
0	6/9/15	DD.MM.YY
1	5/13/15	DD.MM.YY
2	3/18/15	DD.MM.YY
3	12/15/15	MM/DD/YY
4	4/1/15	DD.MM.YY
...
10861	6/15/66	DD.MM.YY
10862	12/21/66	MM/DD/YY
10863	1/1/66	DD.MM.YY
10864	11/2/66	DD.MM.YY
10865	11/15/66	MM/DD/YY

[10866 rows x 2 columns]

Unique date formats: ['DD.MM.YY' 'MM/DD/YY']

```
formats = {  
    r'^\d{4}-\d{2}-\d{2}$': 'YYYY-MM-DD',  
    r'^\d{2}/\d{2}/\d{4}$': 'MM/DD/YYYY',  
    r'^\d{2}/\d{2}/\d{2}$': 'MM/DD/YY',  
    r'^\d{1,2}\s(w+)\s\d{4}$': 'DD Month YYYY',  
    r'^\d{1,2}.\d{1,2}.\d{4}$': 'DD.MM.YYYY',  
    r'^\d{4}.\d{1,2}.\d{1,2}$': 'YYYY.MM.DD',  
    r'^\d{1,2}.\d{1,2}.\d{2}$': 'DD.MM.YY',  
    r'^\s(w+)\s\d{1,2},\s\d{4}$': 'Month DD, YYYY',  
    r'^\s(w+)\s\d{1,2}th,\s\d{4}$': 'Month DDth, YYYY',  
}
```

	release_date	date_format
0	2015-06-09	YYYY-MM-DD
1	2015-05-13	YYYY-MM-DD
2	2015-03-18	YYYY-MM-DD
3	2015-12-15	YYYY-MM-DD
4	2015-04-01	YYYY-MM-DD
...
10861	2066-06-15	YYYY-MM-DD
10862	2066-12-21	YYYY-MM-DD
10863	2066-01-01	YYYY-MM-DD
10864	2066-11-02	YYYY-MM-DD
10865	2066-11-15	YYYY-MM-DD

[10866 rows x 2 columns]

Unique date formats: ['YYYY-MM-DD']

Replaced them to only
one single type of
Date-Time format

genres	production_companies
Action Adventure Science Fiction Thriller	Universal Studios Amblin Entertainment Legenda...
Action Adventure Science Fiction Thriller	Village Roadshow Pictures Kennedy Miller Produ...
Adventure Science Fiction Thriller	Summit Entertainment Mandeville Films Red Wago...
Action Adventure Science Fiction Fantasy	Lucasfilm Truenorth Productions Bad Robot

Values Separated by “|”

Total 5 columns of them
Replaced them with “,”

genres	production_companies
ure, Science Fiction, Thriller	Universal Studios, Amblin Entertainment, Legend...
ure, Science Fiction, Thriller	Village Roadshow Pictures, Kennedy Miller Productions
ence Fiction, Thriller	Summit Entertainment, Mandeville Films, Red Wagon
ure, Science Fiction, Fantasy	Lucasfilm, Truenorth Productions, Bad Robot

```

movies_df['genres'] = movies_df['genres'].apply(lambda g: g.replace('|', ','))
movies_df['production_companies'] = movies_df['production_companies'].apply(lambda x: x.replace('|', ','))
movies_df['cast'] = movies_df['cast'].apply(lambda c: c.replace('|', ','))
movies_df['director'] = movies_df['director'].apply(lambda d: d.replace('|', ','))
movies_df['keywords'] = movies_df['keywords'].apply(lambda k: k.replace('|', ','))
comma_sep = movies_df.loc[:, ['genres', 'production_companies', 'cast', 'director']]
comma_sep.head()

```

```
movies_df['popularity'].head()
```

```
0    32.985763
1    28.419936
2    13.112507
3    11.173104
4     9.335014
```

Popularity being Non-Normalized

```
movies_df['popularity']=1+9*(movies_df['popularity'] -
                             movies_df['popularity'].min())/(movies_df['popularity'].max()-movies_df['popularity'].min())
movies_df['popularity'].head()
```

Pythor

```
0    10.000000
1     8.754235
2     4.577671
3     4.048514
4     3.546999
```

```
Name: popularity, dtype: float64
```

erent
aracters
olumn
pronounce and even

QUESTION


```
def clean_encoded_name(encoded_name):
    cleaned_name = encoded_name.replace("Ã%", "ü") \
        .replace("Ã...Ã¡", "s") \
        .replace("Ã€", "Ãe") \
        .replace("ÃfÃ©", "é") \
        .replace("ÃfÃª", "ú") \
        .replace("ÃfÃ¬", "á") \
        .replace("ÃfÃ±", "ñ") \
        .replace("ÃfÃ", "ó") \
        .replace("ÃfÃ¬", "á") \
        .replace("ÃfÃª", "ú") \
        .replace("ÃfÃ¬", "á") \
        .replace("ÃfÃ%", "ý") \
        .replace("ÃfÃ±", "ä") \
        .replace("Ã¬", "á") \
        .replace("Ã", "í") \
        .replace("Ã¬", "á") \
        .replace("Ã©", "é") \
        .replace("Ãª", "ó") \
        .replace("Ãª", "é") \
        .replace("Ã-", "ö") \
        .replace("Ã±", "ç") \
        .replace("Ã¥", "g") \
        .replace("Ãª", "r") \
        .replace("Ã", "í") \
        .replace("Ãª", "ó")
```

Replacement taken from :

Google (similar readable english letter)

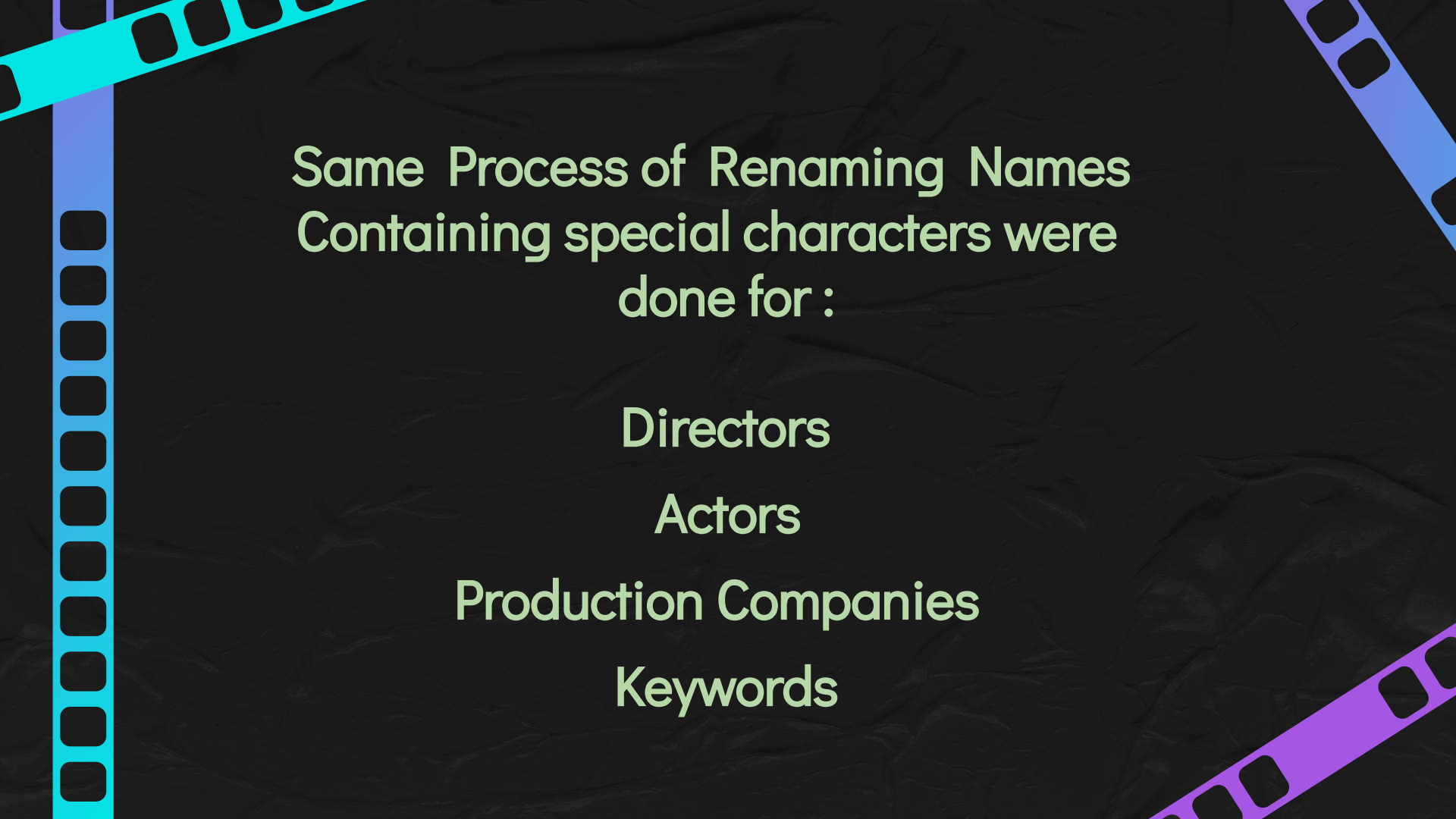
WikiPedia (similar readable english letter)

IMDB (actual english readable names of directors)

```
encoded_names = [
```

1

Alejandro González Iñárritu
Kryštof Hádek
Aki Kaurismäki
José Zúñiga
Cheung Yam-Yim
Alex Brendemühl
Tómas Lemarquis
Irene Montalá
Félix Gómez
Étienne Chatiliez
Çağan Irmak
Jiří Menzel



Same Process of Renaming Names
Containing special characters were
done for :

Directors

Actors

Production Companies

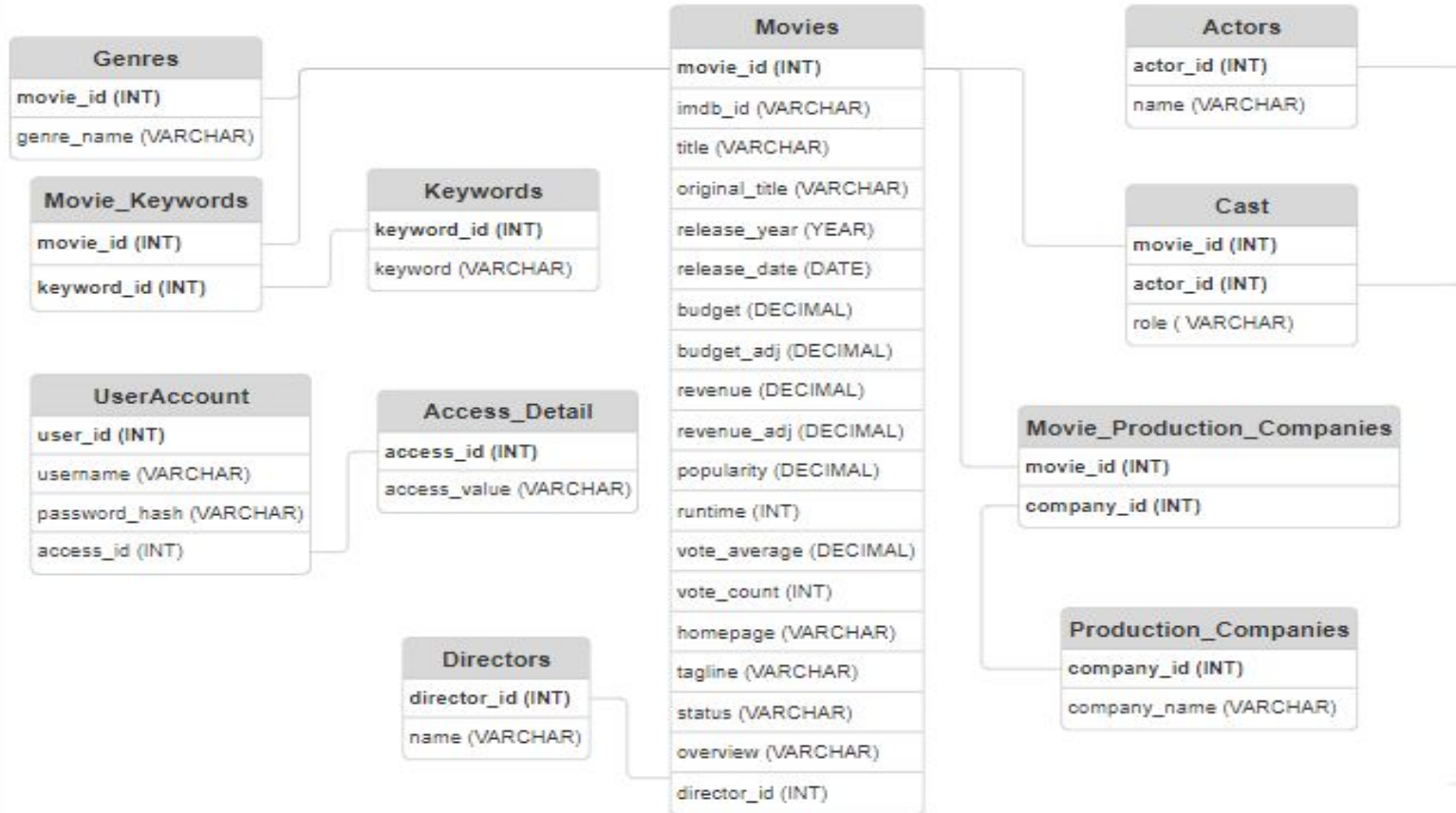
Keywords

DataBase Schema Creation

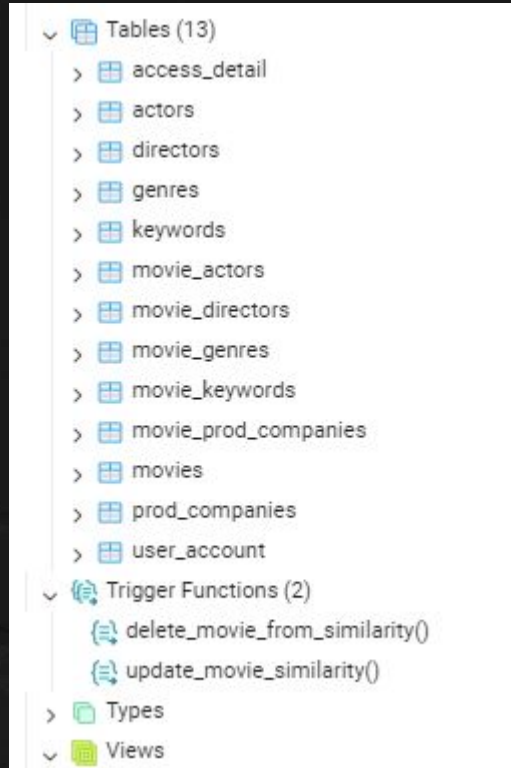
SQLAlchemy
psycopg2

	id	imdb_id	popularity	budget	revenue	original_title
0	135397	tt0369610	32.985763	150000000	1513528810	Jurassic World
1	76341	tt1392190	28.419936	150000000	378436354	Mad Max: Fury Road
2	262500	tt2908446	13.112507	110000000	295238201	Insurgent
3	140607	tt2488496	11.173104	200000000	2068178225	Star Wars: The Force Awakens

DB DESIGN



DB DESIGN



11 Tables for Movie
DataBase

2 Tables For User
Account and Access
Details

DB DESIGN

11 Tables for Movie DataBase

Actors , Directors , Genres, Keywords, Production Companies

-> These all 5 had multiple values inside one column .

Hence , separated them out with new each new table for each
unique name = 5 tables

And For each of them another table to represent their
relationship with the movie ID = 5 tables

One last table containing all the remaining non-redundant
columns (movie table) = 1 Table

DB DESIGN

Table (movie_prod_companies) : [movie_id , prod_comp_id]

Table (prod_companies) : [ID , Name]

Table (movie_directors) : [movie_id , director_id]

Table (directors) : [ID , Name]

Table (movie_genres) : [movie_id , genre_id]

Table (genres) : [ID , Name]

Table (movie_actors) : [movie_id , actor_id]

Table (actors) : [ID , Name]

Table (movie_keywords) : [movie_id , keyword_id]

Table (keywords) : [ID , Name]

Table (movies) : ['id','imdb_id','popularity','budget','revenue','original_title','homepage',
'tagline','overview','runtime','release_date','vote_count','vote_average',
'release_year','budget_adj','revenue_adj']

DB DESIGN

Made appropriate
DataFrames and
directly send them to
database as creating
tables with complete
data

```
table_name = 'keywords'
keyword_df.to_sql(table_name, engine, if_exists='replace', index=False)

print(f"Data successfully inserted into the '{table_name}' table in the '{db_name}' database!")
```

8]

Data successfully inserted into the 'keywords' table in the 'movies_de_database' database!

```
table_name = 'movie_keywords'
movie_keyword_df.to_sql(table_name, engine, if_exists='replace', index=False)

print(f"Data successfully inserted into the '{table_name}' table in the '{db_name}' database!")
```

9]

Data successfully inserted into the 'movie_keywords' table in the 'movies_de_database' database!

```
table_name = 'actors'
actor_df.to_sql(table_name, engine, if_exists='replace', index=False)

print(f"Data successfully inserted into the '{table_name}' table in the '{db_name}' database!")
```

10]

Data successfully inserted into the 'actors' table in the 'movies_de_database' database!

VIEW (REMOVED)

```
Query Query History
1 CREATE EXTENSION IF NOT EXISTS pg_trgm;
2
3 -- DROP MATERIALIZED VIEW IF EXISTS movie_similarity;
4
5
6 ▾ CREATE MATERIALIZED VIEW IF NOT EXISTS movie_similarity AS
7 SELECT
8     LEAST(m1.id, m2.id) AS movie_id_1,
9     GREATEST(m1.id, m2.id) AS movie_id_2,
10     similarity(m1.original_title, m2.original_title) AS similarity_score
11 FROM movies m1
12 JOIN movies m2 ON m1.id != m2.id
13 WHERE similarity(m1.original_title, m2.original_title) > 0.5
14 AND similarity(m1.original_title, m2.original_title) < 1;
15
16
```

```
23
24 ▾ CREATE TABLE IF NOT EXISTS movie_similarity (
25     movie_id_1 INT,
26     movie_id_2 INT,
27     similarity_score REAL,
28     PRIMARY KEY (movie_id_1, movie_id_2)
29 );
30
```


TRIGGER

Query Query History

```
1  CREATE OR REPLACE FUNCTION delete_movie_from_similarity()  
2  RETURNS TRIGGER AS $$  
3  BEGIN  
4      DELETE FROM movie_similarity  
5          WHERE movie_id_1 = OLD.id OR movie_id_2 = OLD.id;  
6  
7      RETURN OLD;  
8  END;  
9  $$ LANGUAGE plpgsql;  
10  
11  
12  CREATE OR REPLACE TRIGGER after_movie_delete  
13  AFTER DELETE ON movies  
14  FOR EACH ROW  
15  EXECUTE FUNCTION delete_movie_from_similarity();  
16  
17  
18
```

Query Query History

```
1  CREATE OR REPLACE FUNCTION update_movie_similarity()  
2  RETURNS TRIGGER AS $$  
3  BEGIN  
4      INSERT INTO movie_similarity (movie_id_1, movie_id_2, similarity_score)  
5      SELECT  
6          LEAST(NEW.id, m.id) AS movie_id_1,  
7          GREATEST(NEW.id, m.id) AS movie_id_2,  
8          similarity(NEW.original_title, m.original_title) AS similarity_score  
9      FROM movies m  
10     WHERE NEW.id != m.id  
11         AND similarity(NEW.original_title, m.original_title) > 0.5  
12         AND similarity(NEW.original_title, m.original_title) < 1;  
13  
14     RETURN NEW;  
15  END;  
16  $$ LANGUAGE plpgsql;  
17  
18  
19  CREATE OR REPLACE TRIGGER after_movie_insert  
20  AFTER INSERT ON movies  
21  FOR EACH ROW  
22  EXECUTE FUNCTION update_movie_similarity();  
23
```

PART - 2

(SUDHIN S)



Implementation

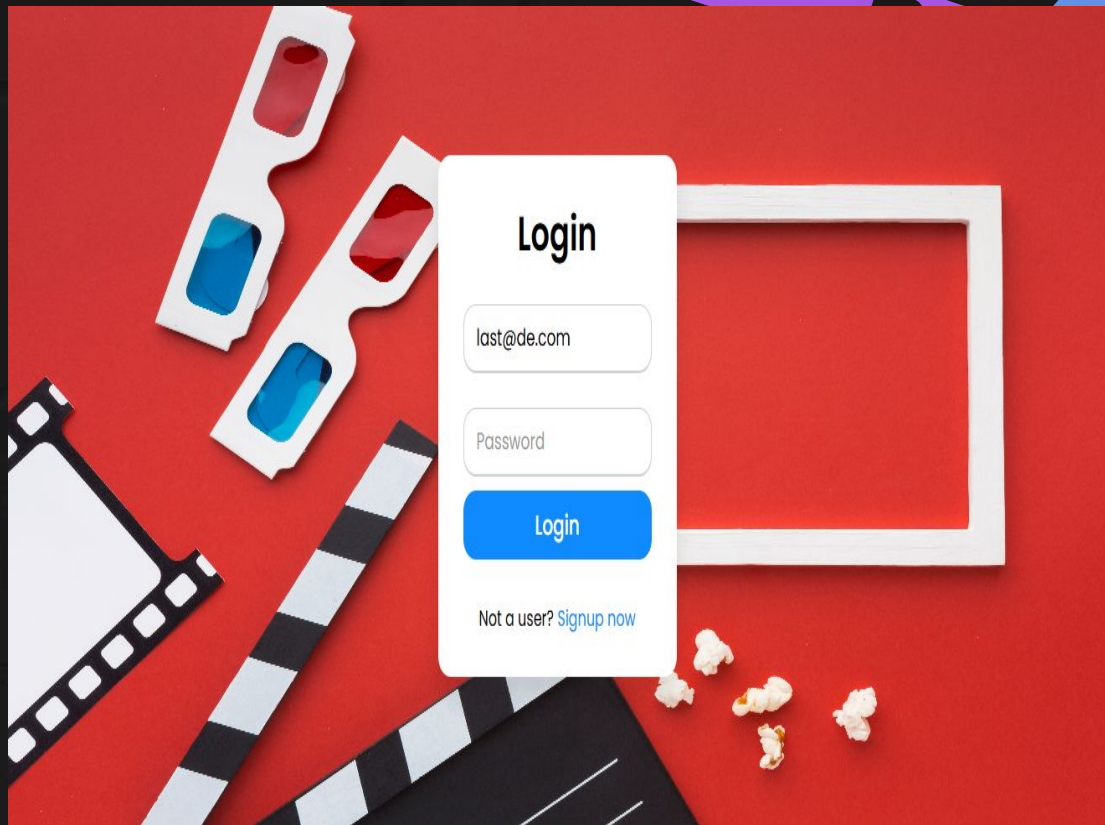
Flask

DASH

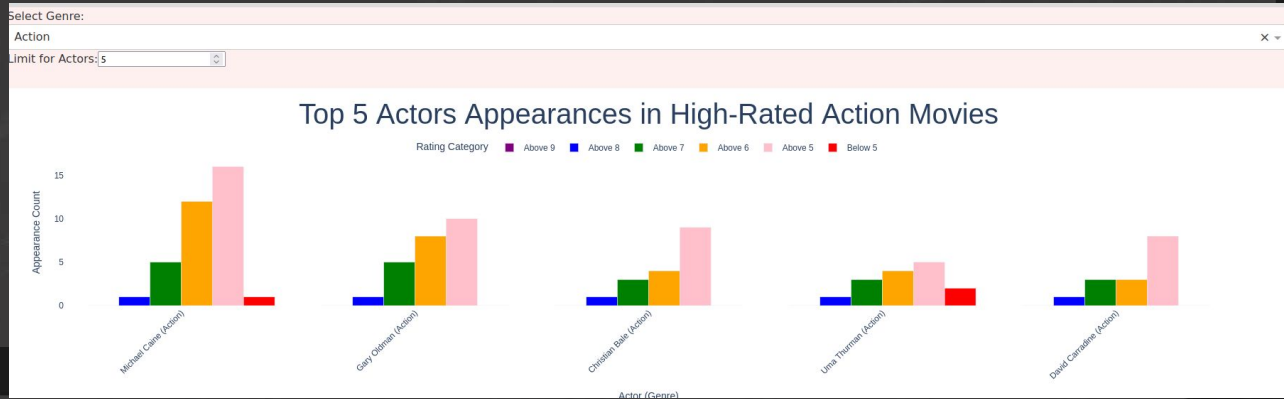
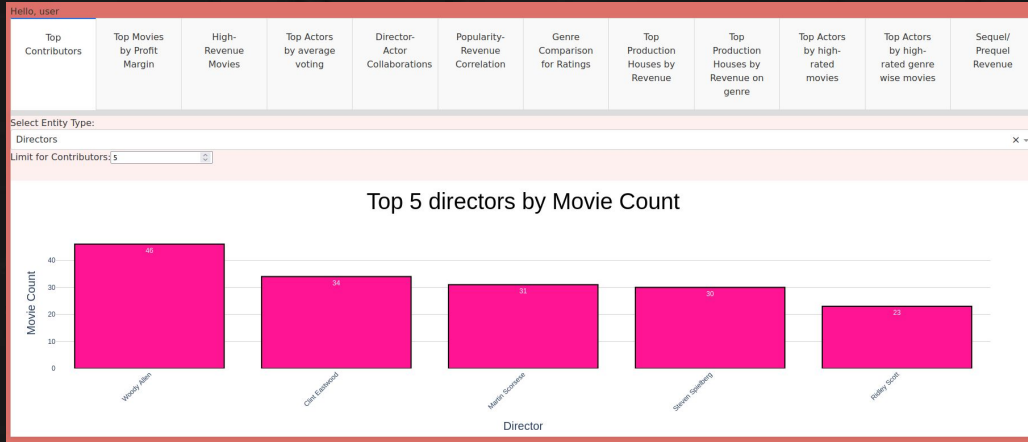
PlotLY

PSYCOPG2

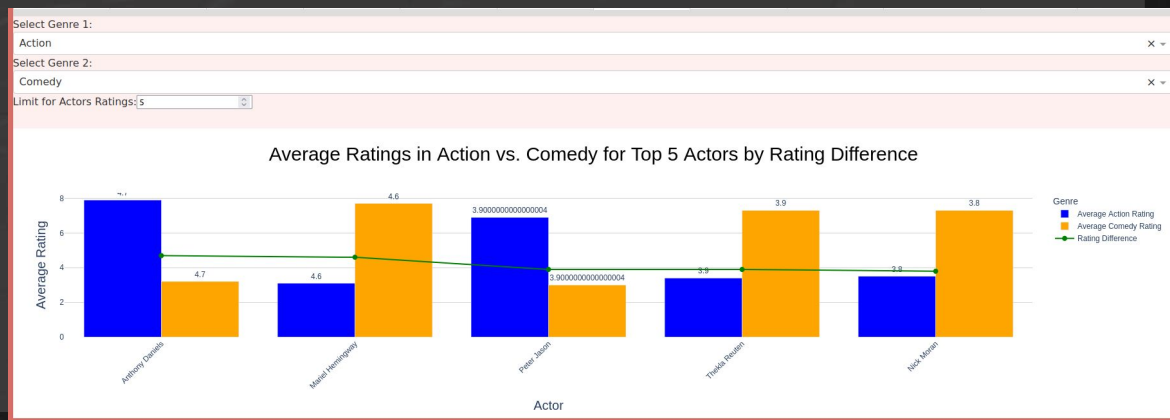
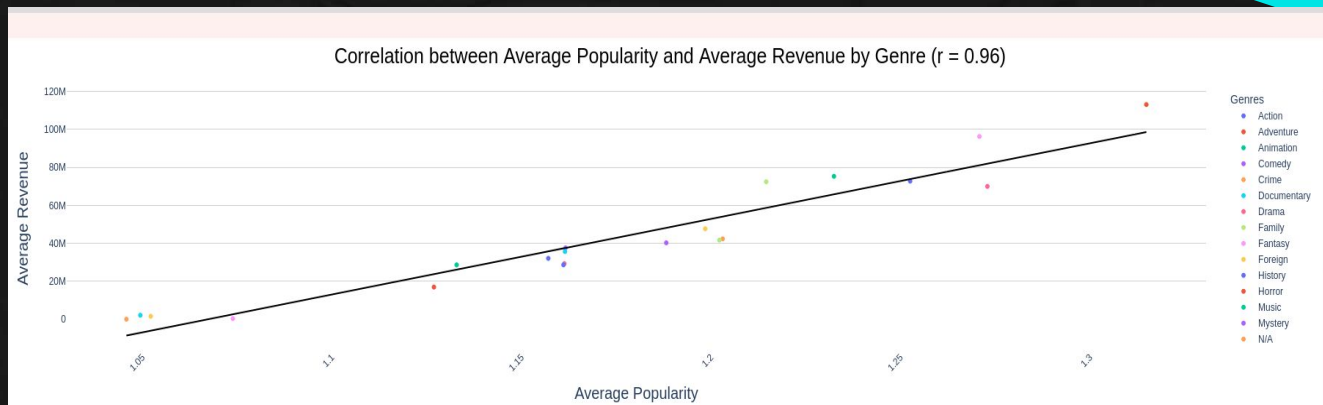
SQLAlchemy



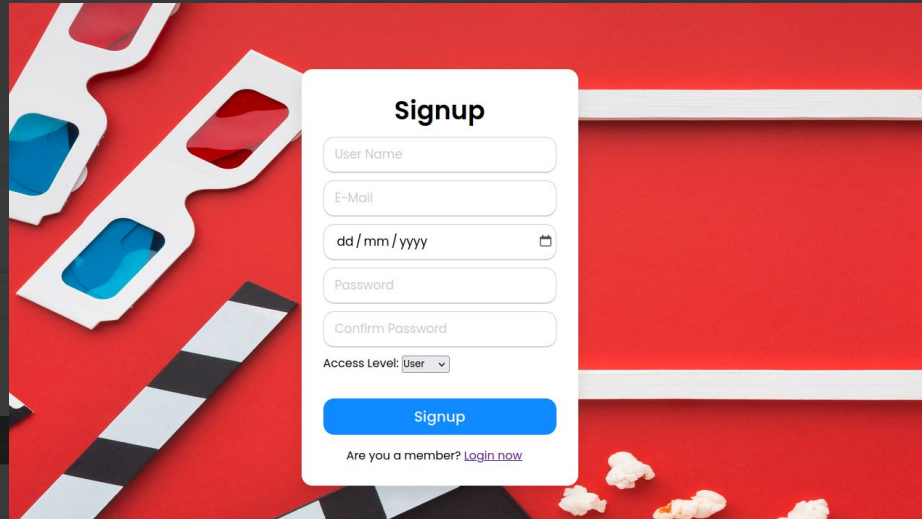
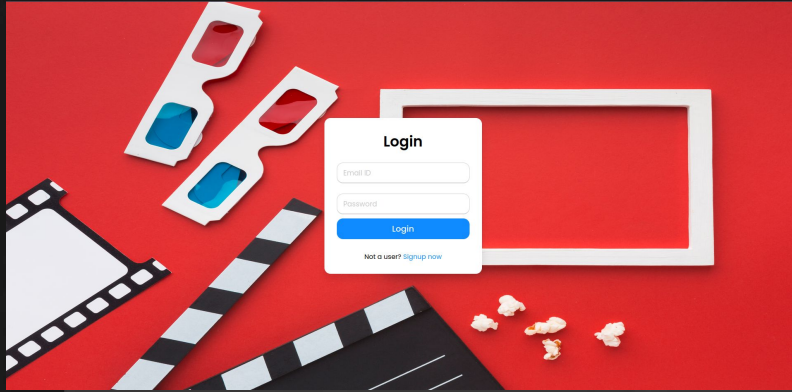
Data Visualizations



Data Visualizations



Login and Signup



Admin Dashboard

Admin Dashboard

[Add New Movie](#)

ID	IMDB ID	Title	Popularity	Budget	Revenue	Release Date	Average Vote	Actions
135397	tt0369610	Jurassic World	10.0	150000000	1513528810	2015-06-09	6.5	Edit Delete
76341	tt1392190	Mad Max: Fury Road	8.75423454734837	150000000	378436354	2015-05-13	7.1	Edit Delete
262500	tt2908446	Insurgent	4.577671086420546	110000000	295238201	2015-03-18	6.3	Edit Delete
140607	tt2488496	Star Wars: The Force Awakens	4.048513661890678	200000000	2068178225	2015-12-15	7.5	Edit Delete
168259	tt2820852	Furious 7	3.546999035763924	190000000	1506249360	2015-04-01	7.3	Edit Delete
281957	tt1663202	The Revenant	3.4857959652695545	135000000	532950503	2015-12-25	7.2	Edit Delete
87101	tt1340138	Terminator Genisys	3.3612853667671363	155000000	440603537	2015-06-23	5.8	Edit Delete
286217	tt3659388	The Martian	3.0919980229007127	100000000	595380321	2015-09-30	7.6	Edit Delete
211672	tt2936440	Minions	3.020175331832005	74000000	1156730962	2015-06-17	6.5	Edit Delete
150540	tt2096673	Inside Out	2.726222407056537	175000000	853708609	2015-06-09	8.0	Edit Delete
206647	tt2379713	Spectre	2.691701445881182	245000000	880674609	2015-10-26	6.2	Edit Delete
76757	tt1617661	Jupiter Ascending	2.6887238826960704	176000003	183987723	2015-02-04	5.2	Edit Delete
264660	tt0470752	Ex Machina	2.669482270770805	15000000	36869414	2015-01-21	7.6	Edit Delete
257344	tt1201020	Pixels	2.632961351916822	88000000	243637091	2015-07-16	5.8	Edit Delete
99861	tt2395427	Avengers: Age of Ultron	2.622028977528382	280000000	1405035767	2015-04-22	7.4	Edit Delete
273248	tt3460252	The Hateful Eight	2.609334293911258	44000000	155760117	2015-12-25	7.4	Edit Delete
199818	tt2236054	Mystery Road	1.1064594115910478	0	0	2013-06-05	6.1	Edit Delete
260346	tt2446042	Taken 3	2.568777989782117	48000000	325771424	2015-01-01	6.1	Edit Delete
102899	tt0478970	Ant-Man	2.5206005645234493	130000000	518602163	2015-07-14	7.0	Edit Delete
150689	tt1661199	Cinderella	2.51613517470511	95000000	542351353	2015-03-12	6.8	Edit Delete
131634	tt1051266	The Hunger Games: Mockingjay - Part 2	2.494345731292392	160000000	650523427	2015-11-18	6.5	Edit Delete
158852	tt1964418	Tomorrowland	2.4903021606515647	190000000	609035668	2015-05-19	6.2	Edit Delete
307081	tt1798684	Southpaw	2.4561762797925333	30000000	91709827	2015-06-15	7.3	Edit Delete
234128	tt2126335	San Andreas	2.339062250554771	110000000	470490832	2015-05-27	6.1	Edit Delete
216015	tt2322441	Fifty Shades of Grey	2.2851943590825305	40000000	569651467	2015-02-11	5.3	Edit Delete

Edit Movie

ID:

135397

IMDB ID:

tt0369610

Original Title:

Jurassic World

Popularity:

10.0

Budget:

150000000

Revenue:

1513528810

Homepage:

<http://www.jurassicworld.com/>

Tagline:

The park is open.

Overview:

Twenty-two years after the events of Jurassic Park, Isla Nublar now features a fully functioning dinosaur theme park, Jurassic World, as originally envisioned by John Hammond.

Runtime (in minutes):

124

Release Date:

09 / 06 / 2015

Vote Count:



Contributions

Shivendr - Pre-processing, Schema Design

Sudhin - Visualization, UI, Authentication

Shivendr & Sudhin - SQL Query , Triggers



Thank You